Stopword Dinamis dengan Pendekatan Statistik

Mardi Siswo Utomo
Program Studi Teknik Informatika Universitas Stikubank Semarang
email:mardiutomo@gmail.com

Abstract *– Stopword a fraction words that often appear in each document corpus. Those words do not have significant meaning for the document. The occurrence of these words make poor index and the retrieval becomes inaccurate. Stopword list or commonly called the stoplist be the most important part in the process of eliminating stopword filtering. Stoplist can be generate from a dictionary or from some references research that generates retrieval stopword list [1]. Stopword depends on the corpus language, so the language provided by stoplist should be the same as the language used in the corpus. Corpus which consists of a variety of languages can not rely on such research stoplist static tuning, Especially if the corpus developed into more than one language and or domain [2].Some words that not include in general stopword could be a stopword inspecific domain corpus. For example the word "recipe" would be a stopword in recipes domain corpus.*

Keyword : *dynamic stoplist, frequency distribution*

Abstrak – Stopword merupakan sebagian kecil kata yang sering muncil pada setiap dokumen korpus. Kata-kata tersebut tidak memberikan makna berarti pada dokumen, sehingga kemunculan kata-kata tersebut dalam indek membuat hasil temu kembali menjadi tidak akurat. Daftar stopword atau biasa disebut dengan stoplist menjadi bagian terpenting dalam proses filtering menghilangkan stopword dari indek temu kembali informasi. Stoplist bisa di dapatkan dari kamus bahasa atau dari beberapa referensi penelitian temu kembali yang menghasilkan daftar stopword [1]. Stopword sangat tergantung dengan bahasa yang digunakan di korpus, sehingga bahasa yang disediakan oleh stoplist harus sama dengan bahasa yang digunakan di korpus. Korpus yang terdiri dari bermacam-macam bahasa tidak bisa mengandalkan stoplist statis seperti pada penelitian tala, Terlebih apabila korpus tersebut berkembang menjadi lebih dari satu bahasa dan atau domain [2]. Demikian pula pada korpus-korpus pada domain yang lebih spesifik beberapa kata yang bukan stopword pada korpus kebanyakan bisa jadi menjadi stopword pada suatu domain korpus. Sebagai contoh kata "resep" akan menjadi stopword pada korpus dengan domain resep masakan.

Kata kunci : stoplist dinamis, distribusi frekuensi

PENDAHULUAN

Perkembangan teknologi internet yang pesat membuat semakin banyaknya pilihan informasi yang tersedia. Terlebih aplikasi berbasis web merupakan aplikasi yang cukup banyak digunakan sekarang ini karena kemudahan dalam penggunaan, implementasi dan perawatan. Perkembangan ini membuat semakin banya informasi yang tersedia di internet tetapi hanya sedikit

informasi yang sesuai dengan keinginan pengguna, selebihnya adalah informasi sampah. Sistem pencarian dan penelusuran informasi yang sesuai menjadi hal penting karena dapat menghemat waktu temu-kembali informasi.

Hampir setiap aplikasi termasuk aplikasi berbasis web dengan pengelolaan basis data membutuhkan proses temu kembali informasi. Pada proses temu kembali selain query dan umpan balik pengguna, terlebih dahulu akan dilakukan pengindekan pada dokumen yang ada. Proses pengindekan data berbasis teks membutuhkan proses filtering pembuangan stopword.

Stopword sendiri adalah Sebagian kecil kata dalam suatu korpus mempunyai jumlah yang sangat berbeda dengan kebanyakan kata lainnya.Sebagai contoh adalah kata "DAN","ITU","INI","DARI" dan "KE", kata-kata tersebut ditemukan hampir disetiap kalimat pada korpus berbahasa Indonesia. Kata-kata semacam ini membuat indek yang terbangun menjadi jelek [2]. Penggunapun tidak mungkin untuk meminta dokumen dengan istilah-istilah ini. Kata-kata ini juga terdapat hampir di semua dan atau paling tidak ada di sebagian besar dokumen. Menurut Francis dan Kucera [3] di dalam sebuah dokumen biasanya terdapat sekitar 20 sampai 30 persen token dari sepuluh kata yang paling sering terjadi dalam bahasa Inggris, Kata-kata ini dikatakan memiliki nilai diskriminasi yang sangat rendah ketika datang ke IR dan mereka dikenal sebagai stopwords dan daftar yang memuat stopword biasa disebut dengan stoplist.

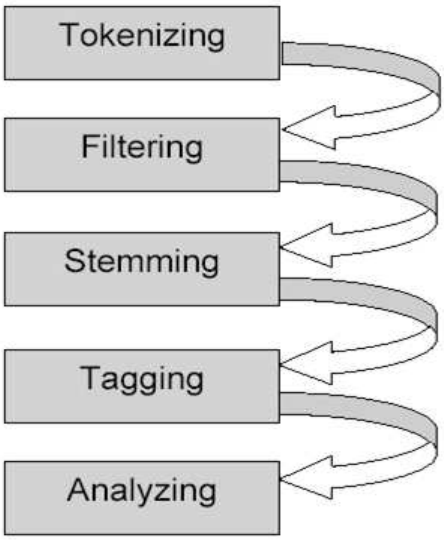
Pada analisa temu kembali informasi stopword merupakan bagian dari informasi yang tidak bermakna, seperti halnya imbuhan. Sehingga stopword harus dihilangkan untuk mempercepat proses pengindekan dan proses query. Proses pembuangan stopword dilakukan dengan acuan kamus / basis data. Stopword yang akan dibuang terlebih dahulu disimpan didalam basis data. Untuk selanjutnya setiap kata pada korpus yang terdapat di dalam basis data stopword akan dibuang oleh sistem. Basis data stopword sendiri sudah banyak tersedia diantaranya oleh Tala [1] ataupun yang disediakan oleh mesin pencari google.Pada korpus yang spesifik seringkali terdapat kata-kata yang tidak bermakna dalam korpus tersebut tetapi tidak terdapat dalam basis data stopword yang ada. Atau sebaliknya kata yang bermakna di korpus tersebut dianggap sebagai stopword dari basis data yang ada.

Pada tulisan ini menyajikan pendekatan statistik dalam menentukan stopword secara dinamis untuk keperluan filtering sistem temu kembali informasi. Korpus yang digunakan pada tulisan ini adalah domain resep masakan Indonesia. Data resep yang digunakan untuk korpus merupakan data resep yang diambil dari 3 Situs resep masakan Indonesia bereputasi yaitu: bango.co.id, masakbagus.com, dan sajiansedap.com. Dari ketiga situs tersebut didapatkan 690 judul resep masakan Indonesia yang akan diproses. Dalam menentukan stopword digunakan pendekatan statistik, pada tulisan ini kami menggunakan distribusi frekuensi dengan batas nilai tertentu untuk menentukan suatu kata merupakan stopword atau tidak. Pada pengujian dilakukan dengan menggunakan algoritma RAKE (Rapid Automatic Keyword Extraction) [4].

TEXT MINING

Text Mining merupakan bagian dari data mining yang mempunyai arti diantaranya adalah : Proses pencarian informasi yang berharga dari sekumpulan data berukuran besar. Data mining juga di artikan sebagai ekplorasi serta analisa data

ukuran besar untuk menemukan pola-pola dan aturan-aturan yang bermakna. Datamining juga dapat didefinisikan dengan sederhana yaitu: mengekstrak dan menambang pengetahuan yang bermanfaat dari data berukuran besar [5].



Gambar 1.Tahapan Penambangan Teks

Menurut Salton [6] tipe informasi dapat dikategorikan menjadi 3 macam yaitu informasi berformat teks, informasi berformat suara dan informasi berformat grafik ataupun gambar. Text mining atau sering disebut text data mining dalam bahasa Indonesia disebut dengan penambangan data teks merupakan proses penambangan data berformat teks dari suatu dokumen. Pada gambar 1 diperlihatkan tahapan-tahapan yang umum dilakukan pada saat melakukan penambangan teks. Proses penambangan melibatkan 5 proses yaitu : a) Tokenizing; b) Filtering; c) Stemming; d) Tagging; e) Analyzing

2.1 Tokenizing

Proses tokenizing adalah proses pemotongan string masukan berdasarkan tiap kata yang menyusunnya. Pada prinsipnya proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenizing mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata. Setelah melalui proses tokenizing maka kalimat tersebut menjadi sekumpulan array yang setiap selnya berisi kata-kata yang ada pada kalimat tersebut. Pada proses tokenizing biasanya juga ditambahkan informasi jumlah kemunculan setiap kata pada kalimat tersebut.

2.2 Filtering

Proses Filtering adalah proses pengambilan kata-kata yang dianggap penting atau mempunyai makna saja. Pada proses ini kata-kata yang dianggap tidak mempunyai makna seperti kata sambung akan dihilangkan. Pada proses ini biasanya digunakan Stop Word List yang tersimpan dalam suatu tabel basis data, yang nantinya digunakan sebagai acuan penghilangan kata. Stop word list berbeda untuk setiap bahasanya. Kata seperti ’di’, ’adalah’ dan ’sebuah’ melalui proses penghilangan, karena kata-kata tersebut tidak mempunyai makna dan hanya berfungsi sebagai kata sambung saja.

2.3 Stemming

Proses stemming adalah proses untuk mencari root dari kata hasil dari proses filtering. Pencarian root sebuah kata atau biasa disebut degan kata dasar dapat memperkecil hasil indeks tanpa harus menghilangkan makna. Filtering adalah proses pengambilan kata-kata yang dianggap penting atau mempunyai makna. Ada dua pendekatan pada proses stemming yaitu pendekatan kamus dan pendekatan aturan. Beberapa penelitian juga telah dilakukan

untuk stemmer bahasa Indonesia baik untuk pendekatan kamus ataupun pendekatan aturan murni adalah Vega [7] dan Tala [1] mereka masing-masing mempunyai algoritma yang berbeda dalam melakukan proses stemmer pada dokumen berbahasa Indonesia.

2.4 Tagging

Proses tagging adalah mencari bentuk utama/root dari suatu kata lampau. Proses tagging tidak digunakan pada dokumen berbahasa Indonesia dikarenakan bahasa Indonesia tidak mengenal kata bentuk lampau.

2.5 Stopword Statis

Merupakan stopword yang sudah ditentukan dari awal sebelum dilakukan proses indek dan filtering pada suatu sistem temu kembali Informasi. Stopword ini biasanya dihasilkan berdasarkan pada telaah kamus bahasa atau dari hasil penelitian sebelumnya. Beberapa institusi penyedia layanan temu kembali infromasi seperti google dan bing juga menyediakan stopword yang mereka gunakan dalam melakukan proses filtering.

Kelebihan dari stopword statis adalah proses filtering stopword menjadi lebih singkat karena tidak ada proses untuk menentukan daftar stopword terlebih dahulu. Kemudahan berbagi pakai dengan korpus lain yang menggunakan bahasa yang sama, karena stopword yang digunakan merupakan kata-kata umum dalam bahasa tersebut.

Kekurangan dari korpus jenis ini adalah pada domain-domain tertentu akan terdapat kata-kata yang bukan stopword dikebanyakan domain tetapi menjadi stopword pada domain tersebut, seperti kata "resep" pada domain masakan Indonesia.

2.6 Stopword Dinamis

Adalah stopword yang dihasilkan dari suatu proses tertentu, proses tersebut biasanya melibatkan korpus yang akan digunakan. Kelebihan dari stopword dinamis adalah kemampuannya dalam beradaptasi dengan korpus atau domain yang spesifik, akurasi temu kembali akan lebih baik jika dibanding menggunakan stopword statik. Kelemahannya adalah dibutuhkan proses komputasi tambahan untuk menghasilkan stopword.

2.7 Distribusi Frekuensi

Hasil pengukuran yang diperoleh biasa disebut dengan raw data atau data mentah. dalam data mentah besarnya nilai dan Jumlah hasil pengukuran yang diperoleh biasanya bervariasi. Sangatlah sulit untuk dapat menarik kesimpulan yang berarti hanya dengan mengamati data mentah tersebut. Untuk memperoleh gambaran terbaik dari data mentah tersebut, maka data mentah tersebut perlu di olah terlebih dahulu.

Pada saat melakukan pengolahan data, akan sangat membantu apabila ada data tersebut diatur dengan cara merangkum data tersebut dengan membuat tabel yang berisi daftar nilai data yang berbeda (baik secara individu atau kelompok) bersama dengan frekuensi yang sejenis [8], yang mewakili berapa kali nilai-nilai tersebut terjadi. Daftar sebaran nilai tersebut dinamakan dengan Daftar Frekuensi atau Sebaran Frekuensi atau Distribusi Frekuensi. Distribusi frekuensi adalah daftar nilai data (berupa nilai individual atau kelompok nilai data tertentu) yang disertai dengan nilai frekuensi yang sesuai.

Pengelompokkan data ke dalam beberapa kelas dimaksudkan agar ciri-ciri penting data tersebut dapat terlihat. Daftar frekuensi diharpkan memberikan gambaran khusus tentang bagaimana keragaman data.

Sifat keragaman data sangat penting untuk diketahui, karena dalam pengujian-pengujian statistik selanjutnya sifat keragaman data tersebut harus diperhatikan. Tanpa memperhatikan sifat keragaman data, penarikan suatu kesimpulan pada umumnya menjadi tidak akurat.

Distribusi frekuensi dibuat dengan alasan berikut: kumpulan data yang besar dapat diringkas, dapat diperoleh beberapa gambaran mengenai karakteristik data, danmerupakan dasar dalam pembuatan grafik.

Proses dalam membangun tabel distribusi frekuensi (selanjutnya disebut TDF) dimulai dari mengurutkan data, biasanya diurutkan dari nilai yang paling kecil, tujuannya agar range data diketahui dan mempermudah penghitungan frekuensi tiap kelas. Tentukan range / jangkauan nilai (1).

Range = nilai maks – nilai min **.....(1)**

Tentukan banyak kelas yang diinginkan. Jangan terlalu banyak/sedikit, berkisar antara 5 dan 20, tergantung dari banyak dan sebaran datanya, untuk selanjutnya bisa digunakan aturan Sturges [8]:

1. Jumlah kelas = 1 + 3.3 log n, dimana n = banyaknya data
2. Tentukan panjang/lebar kelas interval (p)
3. Panjang kelas (p) = [rentang]/[banyak kelas]
4. Tentukan nilai ujung bawah kelas interval pertama

IMPLEMENTASI

Pada saat menyusun Tabel Distribusi Frekuensi, dipastikan bahwa kelas tidak tumpang tindih sehingga setiap nilai-nilai pengamatan harus masuk tepat ke dalam satu kelas. Pastikan juga bahwa tidak akan ada data pengamatan yang tertinggal. Menggunakan lebar yang sama untuk semua kelas, meskipun kadang-kadang tidak mungkin untuk menghindari interval terbuka.

Tabel 1. Tabel indek (45000 record)

id	idkata	jumlah
1	1	1
1	2	5
1	3	4
1	4	2
1	5	2
1	6	1
1	7	1
1	8	1
1	9	1
1	10	1
1	11	1

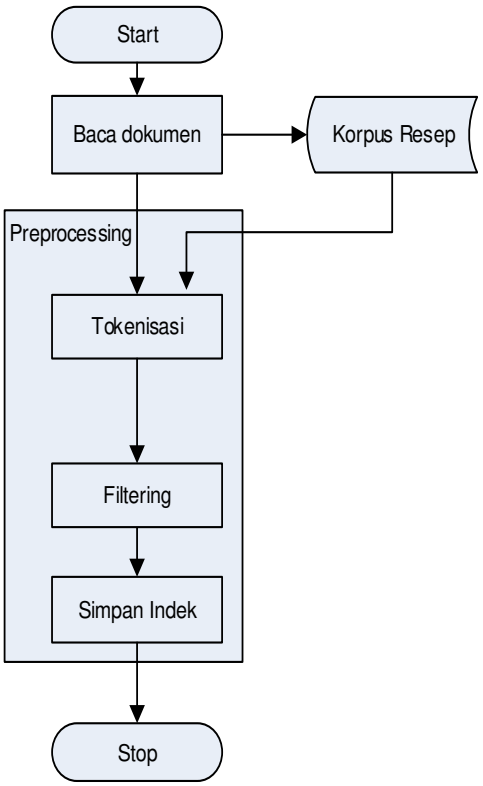
3.1 Proses penyusunan TDF

Sebelum dilakukan proses penyusunan TDF, dilakukan pra poses pada korpus resep masakan Indonesia. Hasil dari proses ini adalah tabel frekuensi kata ataubiasa disebut dengan TF (2) IDF (3)(4) . Contoh hasil tabel pra proses dapat dilihat pada tabel 1, proses ini menghasilkan data mentah sebanyak 44000 record data.

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$ ..... (2)

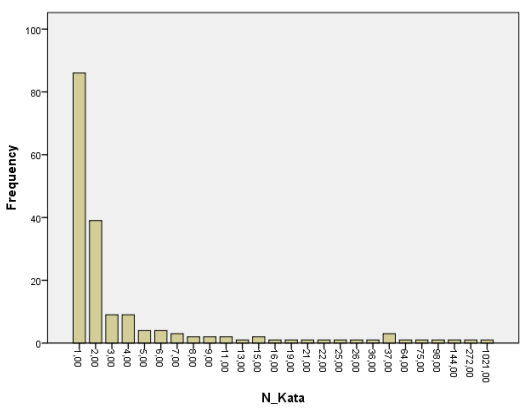
$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$ ..... (3)

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$ ..... (4)



Gambar 2. Diagram alir pra proses

Alur proses pada pra proses di perlihatkan pada gambar 2. Proses diawali dengan mengambil dokumen yang akan diproses, langkah selanjutnya membuang semua artikel tanda baca dan angka pada dokumen tersebut. Proses selanjutnya adalah melakukan proses tokenisasi yang akan memisahkan dokumen menjadi kata-perkata.



Gambar 3. Grafik distribusi kelompok kata

Dari hasil perhitungan didapatkan kelas sebanyak 9 kelas (5) dan interval kelas sebesar 67 (6). Tabel TDF yang dihasilkan diperlihatkan pada tabel 3.

Tabel 2. Tabel kelompok jumlah kata

Idkata	Jumlah
1021	1
272	2
144	3
98	4
64	5
75	6
37	7
37	8
36	9
16	10
37	11
26	12
21	13
25	14
22	15

Alur proses pada pra proses di perlihatkan pada gambar 2. Proses diawali dengan mengambil dokumen yang akan diproses, langkah selanjutnya membuang semua artikel tanda baca dan angka pada dokumen tersebut. Proses selanjutnya adalah melakukan proses tokenisasi yang akan memisahkan dokumen menjadi kata-perkata.

1+(3.3 * log(179)) = 8,43
= 9 .....(5)

598 / 9 = 66,33
= 67 ..... (7)

Tabel 3. Tabel distribusi frekuensi

Kelas	Range	Nilai
1	1-67	2183
2	68-135	65
3	136-203	32
4	204-271	10
5	272-339	13
6	340-407	8
7	408-475	6
8	476-543	9
9	544-611	3

3.2 Pengujian

Setelah didapatkan tabel distribusi frekuensi dari korpus yang ada, filtering menggunakan kata yang termasuk dalam kelas atas (5 kelas terakhir dengan nilai terbesar). Yaitu kelas 272-339,340-407,408-475, 476-543, 544-611. Semua kata anggota dari 5 kelas tersebut ditetapkan sebagai stopword.

Pengujian dilakukan dengan menggunakan algoritma RAKE (Rapid Automatic Keyword Extraction) [7], yang akan mengekstrak kata kunci dari korpus resep masakan Indonesia. Ide dasar dari algoritma RAKE adalah membagi dokumen menjadi kelompok-kelompok kata dengan pemisahan berdasar dari stoplist yang disediakan.

Setiap kelompok kata tersebut dianggap calon kata kunci dan dibobot jumlah kejadian. Rincian metode dapat ditemukan di [3]. Stoplist merupakan hal yang paling penting dan merupakan parameter bebas dari algoritma RAKE, karena merupakan satu-satunya cara untuk menyesuaikan algoritma ini untuk korpus dengan bahasa dan domain yang berbeda.

Gambar 4 merupakan kode program untuk scoring dari algoritma RAKE yang digunakan, sourcecode lengkap dapat diunduh di [9]. Bahasa pemrograman yang digunakan adalah PHP dengan basis data Mysql. Pada tabel 4 diperlihatkan hasil ekstrak kata kunci dari 5 kata kunci dengan bobot terbesar menggunakan algoritma RAKE.

```
private function get_scores($phrases){
  $frequencies = array();
  $degrees = array();
  foreach ($phrases as $p){
   $words = self::split_phrase($p);
   $words_count = count($words);
   $words_degree = $words_count - 1;
   foreach ($words as $w){
    $frequencies[$w] = (isset($frequencies[$w]))? $frequencies[$w] : 0;
    $frequencies[$w] += 1;
    $degrees[$w] = (isset($degrees[$w]))? $degrees[$w] : 0;
    $degrees[$w] += $words_degree;
   }
  }
  foreach ($frequencies as $word => $freq)$degrees[$word] += $freq;
  $scores = array();
  foreach ($frequencies as $word => $freq){
```

```
  $scores[$word] = (isset($scores[$word]))? $scores[$word] : 0;
  $scores[$word] = $degrees[$word] / (float) $freq;
 }
 return $scores;
}
```

Gambar 4. Kode program untuk scoring dari algoritma RAKE

KESIMPULAN DAN SARAN

Pada tabel 5 digunakan stoplist yang disediakan oleh tala [1]. Pada tabel tersebut terlihat bahwa kata kunci yang terekstrak relatif panjang (73 Kata) menjadikan keyword hasil tidak akurat masih mengandung derau kata. Sedangkan pada tabel 5 hasil dari ektraksi keyword menjadi lebih baik dengan panjang keyword terpanjang adalah 10, akurasi dengan kumpulan korpus diharapkan juga semakin baik.

Untuk selanjutnya perlu dilakukan pengukuran akurasi hasil temu kembali dengan menggunakan stopword dinamis, sehingga di peroleh kesimpulan yang lebih baik tentang batasan distribusi frekuensi optimal yang digunakan untuk menentukan stopword dinamis.

Disarankan untuk melakukan kategorisasi secara manual pada korpus yang digunakan dan melakukan evaluasi akurasi hasil temu kembali melalui proses clustering serta mengukur hasilnya dengan precison and recall [10].

Selain itu juga dapat dilakukan pendekatan statistik lainnya seperti menggunakan metode distribusi probabilitas poisson maupun binominal[11], yang dimungkinkan akan menghasilkan akurasi keluaran yang berbeda.

Tabel 4.10 Keyword teratas dengan stoplist umum

id	teks	wcount	nilai
5535	manisnya kecap bango buah tomat cincang kasar buah...	73	4555.33
1207	kecap bango manis pedas gurih terimakasih kecap ba...	66	4099.53
6823	sambal terasi sdm garam sdm gula jawa sdm terasi b...	70	4085.30
681	merebus sendok teh garam sendok teh lada putih bub...	69	3633.97
7176	nikmat batang daun bawang iris halus butir telur a...	66	3348.33
7255	berkurang sdt garam batang daun pandan muda iris l...	64	3156.42
4337	kecap manis bango sdt arak beras sdt bumbu ngohyon...	57	2995.67
5556	sesuai selera iris tipis buah jeruk limo buah toma...	61	2841.98
6968	bagikan resep semur hati ayam pedas buah tomat sen...	59	2829.75
3272	lezaaattt buah bombay sendok makan air asam jawa s...	54	2719.50

Tabel 5. 10 Keyword teratas dengan stoplist umum dan dinamis

id	teks	wcount	nilai
1655	memasukkan citarasa belimbing wuluh nan segar dala...	10	84.20
5036	warung sate klatak mak adi terletak jalan imogiri ...	10	82.13
3989	irisan cabai rawit ditaburi nori rumput laut bungk...	10	74.13
4603	mantab banget deh coba aja kalo percaya ikat kangk...	9	72.70
4005	jgan lupa dibersihkan kepala cumi masukkan kedalam...	9	68.67
3237	jenis makanan kebanggaan khas jawa timur dinamakan...	9	64.80
1129	membuatnyapun mudah temukan resep makanan tradisio...	8	64.00
3302	sandung lamur dipotong kotak kotak didalam wajan p...	8	62.50
1333	kaya citarasa rempah rempah khas kuliner indonesia...	8	62.00
4496	legenda kuliner oseng oseng mercon narti perjalana...	9	60.92

DAFTAR PUSTAKA

[1] Tala, Z, 2003, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands.

[2] Rachel TL, Ben H,Iadh O,2004, Automatically Building a Stopword List for an Information Retrieval System, Department of Computing Science University of Glasgow

[3] W. Francis,1982, Frequency Analysis of English Usage: Lexicon and Grammar Houghton Miffin.

[4] Rose, S., Engel, D., Cramer, N. & W. Cowley, W. (2010). Automatic keyword extraction from individual documents. Text Mining: Applications and Theory. John Wiley & Sons, Ltd.

[5] Han, J dan Kamber, M, 2000, Data Mining : Concept and Techniques, Morgan Kaufmann Publisher.

[6] Salton, G. & Yang, S. (1973). On the specification of term values in automatic indexing, Journal of Documentation

[7] Vega, V.B., (2001), Information Retrieval for the Indonesian Language, Master's thesis, National University of Singapore.

[8] Hasan, M. Iqbal. 2001. Pokok-pokok Materi Statistik I (Statistik Deskriptif), Bumi Aksara. Jakarta.

[9] https://github.com/Richdark/RAKE-PHP/diakses tanggal 10 November 2015

[10] Baesa, R dan Ribeiro, B, 1998, *Modern Information Retrieval*, ACM Press New York USA

[11] Michał J,and Michał Ł,2014,Unsupervised Keyword Extraction From Polish Legal Texts, Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland